CS 7545: Machine Learning Theory **Fall 2021**

Lecture 4

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Estimating Mixtures of Gaussians with Third Moments

Last lecture we saw that second moments were not enough to estimate the parameters of a mixture of Gaussians. Recall that $\mathbf{T} = \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}]$ has $T_{ijk} = \mathbb{E}[x_i x_j x_k]$ and that tensors define a polynomial:

$$\mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \sum_{i,j,k=1}^{d} T_{ijk} u_i v_j w_k. \tag{1}$$

This gives the following expression:

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{x}^\intercal \boldsymbol{v})^3] &= \mathbb{E}\Big[\sum_i x_i v_i \sum_j x_j v_j \sum_k x_k v_k\Big] && (2)\\
&= \sum_{ijk} \mathbb{E}[x_i x_j x_k] v_i v_j v_k && (3)\\
&= \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}](\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}). && (4)
\end{aligned}$$

Let's see if third moments can help us estimate a single Gaussian. Then,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] &= \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu}) \otimes (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu}) \otimes (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu})]\\
&= \boldsymbol{\mu} \otimes \boldsymbol{\mu} \otimes \boldsymbol{\mu}\\
&\quad + \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu}) \otimes (\boldsymbol{x} - \boldsymbol{\mu}) \otimes (\boldsymbol{x} - \boldsymbol{\mu})]\\
&\quad + \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu}) \otimes (\boldsymbol{x} - \boldsymbol{\mu}) \otimes \boldsymbol{\mu}]\\
&\quad + \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu}) \otimes \boldsymbol{\mu} \otimes (\boldsymbol{x} - \boldsymbol{\mu})]\\
&\quad + \mathbb{E}[\boldsymbol{\mu} \otimes (\boldsymbol{x} - \boldsymbol{\mu}) \otimes (\boldsymbol{x} - \boldsymbol{\mu})].
\end{aligned} \tag{5}$$

This expression looks gross, but it's actually not that bad. Let's parse it one term at a time.

1. This is what we will use to help us estimate the mean.

2. This is the third moment, which for a spherical Gaussian is zero.

3. The first two terms are the second moment, so the term is equal to $\sigma^2 \mathbf{I} \otimes \boldsymbol{\mu}$.

4. We'll use a useful identity here: $\mathbf{I} = \sum_{\ell=1}^{d} \boldsymbol{e}_\ell \otimes \boldsymbol{e}_\ell$ where $\boldsymbol{e}_\ell$ is the zero vector with a one in

the ℓ^{th} position. Consider the ijk^{th} position of the term:

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})\otimes\boldsymbol{\mu}\otimes(\boldsymbol{x}-\boldsymbol{\mu})]_{ijk} &= \mathbb{E}[(x_i-\mu_i)\mu_j(x_k-\mu_k)] && (6)\\
&= \mathbb{E}[(x_i-\mu_i)(x_k-\mu_k)\mu_j] && (7)\\
&= \begin{cases} 0 & i\neq k,\\ \sigma^2\mu_j & i=k. \end{cases} && (8)
\end{aligned}$$

So the term is equal to $\sigma^2\sum_{\ell=1}^{d}\boldsymbol{e}_\ell\otimes\boldsymbol{\mu}\otimes\boldsymbol{e}_\ell$.

5. The second two terms are the second moment, so the term is equal to $\boldsymbol{\mu}\otimes\sigma^2\mathbf{I}$.

Using the identity on (3) and (5) we obtain:

Lemma 1.1: Tensor Identity for Mixtures of Gaussians

If $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\mathbf{I})$ then

$$\mathbb{E}[\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}] = \boldsymbol{\mu}\otimes\boldsymbol{\mu}\otimes\boldsymbol{\mu}+\sigma^2\sum_{\ell=1}^{d}\boldsymbol{e}_\ell\otimes\boldsymbol{e}_\ell\otimes\boldsymbol{\mu}+\boldsymbol{e}_\ell\otimes\boldsymbol{\mu}\otimes\boldsymbol{e}_\ell+\boldsymbol{\mu}\otimes\boldsymbol{e}_\ell\otimes\boldsymbol{e}_\ell. \tag{9}$$

Suppose F is a mixture of k Gaussians with weights w_i. Then,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}] &= \sum_{i=1}^{k} w_i\boldsymbol{\mu}_i\otimes\boldsymbol{\mu}_i\otimes\boldsymbol{\mu}_i\\
&\quad+\sum_{i=1}^{k} w_i\sigma_i^2\sum_{\ell=1}^{d}\boldsymbol{e}_\ell\otimes\boldsymbol{e}_\ell\otimes\boldsymbol{\mu}_i+\boldsymbol{e}_\ell\otimes\boldsymbol{\mu}_i\otimes\boldsymbol{e}_\ell+\boldsymbol{\mu}_i\otimes\boldsymbol{e}_\ell\otimes\boldsymbol{e}_\ell.
\end{aligned} \tag{10}$$

The term $\sum_i w_i\sigma_i^2$ is a scalar, so we can attach it to the $\boldsymbol{\mu}_i$. Suppose $\boldsymbol{u}=\sum_i w_i\sigma_i^2\boldsymbol{\mu}_i$, then the formula becomes

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}] &= \sum_{i=1}^{k} w_i\boldsymbol{\mu}_i\otimes\boldsymbol{\mu}_i\otimes\boldsymbol{\mu}_i\\
&\quad+\sum_{\ell=1}^{d}\boldsymbol{e}_\ell\otimes\boldsymbol{e}_\ell\otimes\boldsymbol{u}+\boldsymbol{e}_\ell\otimes\boldsymbol{u}\otimes\boldsymbol{e}_\ell+\boldsymbol{u}\otimes\boldsymbol{e}_\ell\otimes\boldsymbol{e}_\ell.
\end{aligned} \tag{11}$$

This lemma shows that if we are able to estimate $\boldsymbol{u}$, then we could estimate the $\boldsymbol{\mu}_i$. Luckily, this is easy to do with the mean of the mixture. Suppose $\boldsymbol{v}$ is orthogonal to $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$. Then,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}((\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{v})^2] &= \mathbb{E}[\boldsymbol{x} \otimes (\boldsymbol{x}-\boldsymbol{\mu}) \otimes \boldsymbol{x}-\boldsymbol{\mu})](\cdot, \boldsymbol{v}, \boldsymbol{v}) && (12)\\
&= \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}](\cdot, \boldsymbol{v}, \boldsymbol{v}) \\
&\quad + \mathbb{E}[\boldsymbol{x} \otimes -\boldsymbol{\mu} \otimes (\boldsymbol{x}-\boldsymbol{\mu})](\cdot, \boldsymbol{v}, \boldsymbol{v}) && (13)\\
&\quad + \mathbb{E}[\boldsymbol{x} \otimes (\boldsymbol{x}-\boldsymbol{\mu}) \otimes -\boldsymbol{\mu}](\cdot, \boldsymbol{v}, \boldsymbol{v}) \\
&= \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}](\cdot, \boldsymbol{v}, \boldsymbol{v}) && (14)
\end{aligned}$$

By the lemma, this is equal to $\boldsymbol{u}$.

2 Tensor Decomposition

We almost have enough ingredients to estimate the mixture, but we need to figure out how to extract the individual means from $\mathbf{T} = \sum_i w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$. This is called tensor decomposition, and when the $\boldsymbol{\mu}_i$ are orthogonal (which we can make them via a linear transformation) there is a unique solution. This is one reason why second moments did not work but third moments does – the lower-dimensional version of the problem, matrix decomposition, does not result in unique solutions.

We can use a strategy called tensor power iteration to solve this decomposition. We start with a random unit vector $\boldsymbol{x}$, then repeat

$$\boldsymbol{x} \leftarrow \frac{\mathbf{T}(\cdot, \boldsymbol{x}, \boldsymbol{x})}{\|\mathbf{T}(\cdot, \boldsymbol{x}, \boldsymbol{x})\|}. \tag{15}$$

This will converge to some $\boldsymbol{x}^{(t)}$, which one can show is one of the $\boldsymbol{\mu}_i$. Then, we can remove that $\boldsymbol{\mu}_i$ from the sum and repeat.

3 Algorithm for Learning Mixtures of Gaussians

Here is the full algorithm for estimating the mixture parameters. Recall that the only assumption is that the means are linearly independent.

1. Set $\mathbf{M} = \mathbb{E}[x \otimes \boldsymbol{x}]$ and compute the top k eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$. Set $\hat{\sigma}^2$ as the $(k+1)^{st}$ eigenvalue of $\mathbf{M}$.

2. Factorize $(\mathbf{M} - \hat{\sigma}^2 \mathbf{I} = \mathbf{W}\mathbf{W}^\intercal$ and apply $\hat{\mathbf{S}} = \mathbf{W}^{-1}\mathbf{S}$ to make the means orthogonal.

3. Pick $\boldsymbol{v} \perp \{\mathbf{W}^{-1}\boldsymbol{v}_1, \mathbf{W}^{-1}\boldsymbol{v}_2, \ldots, \mathbf{W}^{-1}\boldsymbol{v}_k\}$. Then set $\boldsymbol{u} = \mathbb{E}[\boldsymbol{x}((\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{v})^2]$ and compute

$$\mathbf{T} = \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] - \sum_{\ell=1}^{d} \boldsymbol{e}_\ell \otimes \boldsymbol{e}_\ell \otimes \boldsymbol{u} + \boldsymbol{e}_\ell \otimes \boldsymbol{u} \otimes \boldsymbol{e}_\ell + \boldsymbol{u} \otimes \boldsymbol{e}_\ell \otimes \boldsymbol{e}_\ell. \tag{16}$$

4. Apply tensor power iteration to $\mathbf{T}$ to get $\boldsymbol{y}$. Set

$$\hat{\boldsymbol{\mu}}_i = \mathbf{T}(\boldsymbol{y}, \boldsymbol{y}, \boldsymbol{y})\boldsymbol{y} \tag{17}$$

$$\hat{w}_i = \frac{1}{\|\hat{\boldsymbol{\mu}}_i\|^2} \tag{18}$$

$$\hat{\sigma}_i^2 = \boldsymbol{u}^\intercal \hat{\boldsymbol{\mu}}_i. \tag{19}$$

Note that the main computational difficulty in this algorithm is computing $\mathbf{T}$, but we don't actually have to compute the whole thing. We can just use the vectors we need each time. A final note is that the complexity depends on the condition number of the means – if they are almost linearly dependent, the algorithm could take much longer.